• CORPES: EL ESPAÑOL DEL SIGLO XXI

El Corpus del Español del Siglo XXI (CORPES XXI) forma parte de un proyecto académico panhispánico iniciado en 2007 cuyo objetivo final es reunir, en 2014, un total de 300 millones de formas de la lengua común de 450 millones de hispanohablantes. Este corpus, que supone la continuación de los ya terminados CREA y CORDE, se comenzó a elaborar en 2007, a partir de textos orales y escritos. El material proviene tanto de medios impresos —libros y prensa— como de contenidos publicados en Internet o emitidos en canales de información audiovisual.

El CORPES XXI es una iniciativa de la RAE y de la Asociación de Academias de la Lengua Española (ASALE) en la que trabajan, además, ocho equipos externos a las Academias: seis de diferentes universidades españolas —Alcalá de Henares, Autónoma de Barcelona, León, Salamanca, Santiago de Compostela y Valencia—, la Academia Argentina de Letras y la Fundación Comillas.

Tal como se puso de manifiesto en un informe presentado en el XIV Congreso de la ASALE, celebrado en Panamá a finales de 2011, el CORPES XXI consta actualmente de cerca de 100 millones de formas. La previsión para los próximos tres años es incrementar esa cifra en 200 millones más —hasta llegar a 300—, mediante la correspondiente selección, codificación e integración de materiales.

Los textos que servirán de base al CORPES XXI reflejarán adecuadamente el español de todo el mundo: el 30 % de las formas procederán de España y el 70 % restante de América. Se seleccionarán geográficamente, de acuerdo con una serie de criterios relativos al número de hablantes de cada área lingüística, la publicación de libros y la representación de medios y canales de comunicación en Internet.

El resultado de esta ambiciosa recopilación será la creación de una gran base de datos que, unida a un potente programa de recuperación, permitirá consultar los ejemplos por países, fechas o temas. El CORPES XXI constituirá una herramienta imprescindible para investigadores, lexicógrafos y cualquier estudioso a quien puedan interesar los datos de los últimos años de la lengua española.

LOS ANTECEDENTES: CREA Y CORDE

El CORPES XXI tiene como antecedentes históricos el Corpus de Referencia del Español Actual (CREA) y el Corpus Diacrónico del Español (CORDE), cuya confección fue puesta en marcha por la RAE en la última década del siglo XX. En este conjunto de textos de todas las épocas del español, procedentes de los países de habla hispana y de los más diversos tipos, se encuentran los materiales relevantes que los lexicógrafos y los gramáticos necesitan para llevar a cabo su trabajo.

La unión de ambos corpus [cerca de 300 millones de formas desde los orígenes del idioma hasta 1974 en el CORDE y algo más de 155 millones de formas desde 1975 hasta 2004 en el CREA] proporciona a todos los investigadores o simples interesados en nuestra lengua el recurso en el que poder documentar con comodidad, rapidez y seguridad la mayor o menor frecuencia con que se utiliza una palabra, su distribución por países, años, tipos de texto, áreas temáticas, etc.

En definitiva, estos dos corpus contienen cuanto se necesita para trabajar sobre bases sólidas, tanto en la línea estrictamente científica como en la que fundamenta la toma de decisiones normativas que la Asociación de Academias tiene a su cargo en todo el mundo hispánico.

El nuevo CORPES XXI es concebido como un corpus de referencia en el sentido actual de la expresión y, por tanto, ocupa un lugar intermedio entre los corpus especializados, por un lado, y la utilización directa de la enorme cantidad de materiales contenidos en la Red, por otro. Los corpus especializados pueden estar mucho más cuidados en la codificación de los textos y en la atención filológica que se les presta, pero están enfocados a tipos muy específicos de materiales y, en consecuencia, no pueden cubrir adecuadamente las necesidades generales de la investigación. En cambio, la utilización de los textos existentes en la Red —la tendencia conocida como *web as corpus*— tiene la ventaja de su carácter gratuito y la ingente cantidad de formas que se pueden examinar, pero carece de la posibilidad de hacer recuperación selectiva de la información, que es precisamente lo que necesita la investigación teórica o aplicada.

El Banco Santander respalda esta iniciativa de la Real Academia Española a través de su División Global Santander Universidades, cuyas actividades vertebran la acción social de la entidad bancaria y le permiten mantener una alianza estable con más de 990 universidades e instituciones académicas en América, Asia y Europa. Más información en www.santander.com/universidades.

CORPES XXI
Estado actual, 8. 03. 2012

Número total de formas: 98 762 224

Procedencia:

CREA 2001-2004	27 943 719
CORPES 1ª y 2ª fase	54 464 507
CORPES 3ª fase	16 353 998

TOTAL CORPES (sin CREA): 70 818 505

Distribución CORPES España / América

España	23 992 210
América	46 826 295

Distribución por años

2004	423 428
2005	11 901 560
2006	13 772 433
2007	13 092 084
2008	11 815 995
2009	7 531 747
2010	8 425 559
2011	3 855 699

Distribución por zonas

Andina	7 073 719
Antillas	7 181 554
Caribe continental	7 224 978
Chilena	3 042 998
España	23 992 210
Estados Unidos	2 009 032
México y Centroamérica	11 213 204
Río de la Plata	9 080 810

REAL ACADEMIA ESPAÑOLA

Santander

Distribución por países

Argentina	5 756 521
Bolivia	2 527 745
Chile	3 042 998
Colombia	3 644 395
Costa Rica	995 829
Cuba	3 043 387
Ecuador	2 030 467
El Salvador	481 838
España	23 992 210
Estados Unidos	2 009 032
Guatemala	644 106
Honduras	460 885
México	6 809 402
Nicaragua	1 282 921
Panamá	538 223
Paraguay	1 463 476
Perú	2 515 507
Puerto Rico	1 844 152
República Dominicana	2 294 015
Uruguay	1 860 813
Venezuela	3 580 583

Distribución por bloques

Ficción	5 545 150
No_ficción	65 273 355

Distribución por temas

Ciencias y tecnología	8 221 566
Ciencias sociales, creencias y pensamiento	13 068 976
Política economía y justicia	18 493 639
Artes, cultura, espectáculos	9 442 569
Actualidad, ocio y vida cotidiana	10 317 766
Salud	5 728 492
Novela	5 130 343
Relatos	298 173
Teatro	116 981

Distribución por medio

Libro	12 689 306
Prensa	55 700 566
Web	2 428 633